

# Kwaliteit van klinisch onderwijs

Citation for published version (APA):

Wolfhagen, H. A. P. (1993). *Kwaliteit van klinisch onderwijs*. [Doctoral Thesis, Maastricht University]. Datawyse / Universitaire Pers Maastricht. <https://doi.org/10.26481/dis.19930218hw>

**Document status and date:**

Published: 01/01/1993

**DOI:**

[10.26481/dis.19930218hw](https://doi.org/10.26481/dis.19930218hw)

**Document Version:**

Publisher's PDF, also known as Version of record

**Please check the document version of this publication:**

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

**General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

[www.umlib.nl/taverne-license](http://www.umlib.nl/taverne-license)

**Take down policy**

If you believe that this document breaches copyright please contact us at:

[repository@maastrichtuniversity.nl](mailto:repository@maastrichtuniversity.nl)

providing details and we will investigate your claim.

# SAMENVATTING

Het thema van dit proefschrift is de ontwikkeling van een systeem voor kwaliteitszorg van het klinisch onderwijs, dat wil zeggen de praktisch-medische stages/co-assistentschappen in ziekenhuizen. In *Hoofdstuk 1* staat de context waarbinnen deze studie verricht is centraal en komt de uitwerking van de probleemstelling aan de orde.

Ook binnen het universitaire onderwijs is er een toenemende belangstelling voor de kwaliteit van het onderwijs. Dit komt met name door een aantal maatregelen (HOAK, 1985; HOOP, 1987) van het ministerie van Onderwijs en Wetenschappen waarbij de universiteiten gevraagd wordt verantwoording af te leggen over de kwaliteit van hun onderwijs. Een logisch gevolg van deze vraag naar informatie is dat binnen het hoger onderwijs behoefte was aan systemen die dit mogelijk zouden maken. Aan de Rijksuniversiteit Limburg werd al geruime tijd voordat de kwaliteitszorg van onderwijs in Nederland actueel werd, gewerkt aan de ontwikkeling van systemen voor kwaliteitszorg. De redenen hiervoor zijn dat deze faculteit relatief jong is en een onderwijsstrategie - probleemgestuurd leren - hanteert die afwijkt van die van de zusterfaculteiten. Omdat de benadering van het onderwijs anders was, ontstond binnen haar eerste faculteit te weten die der geneeskunde behoefte aan informatie over de kwaliteit van het onderwijs ter ondersteuning van curriculumontwikkeling. Daarnaast was het doel verantwoording af te leggen aan zusterfaculteiten en de overheid. Aanvankelijk richtte de aandacht zich voornamelijk op de eerste vier studiejaar, maar sedert 1987 ook op het vijfde en zesde studiejaar. Gedurende deze twee laatste studiejaar lopen studenten stages binnen ziekenhuizen, huisartspraktijken en instellingen van de gezondheidszorg, de zogenaamde "co-assistentschappen". De problematiek die in dit proefschrift aan de orde wordt gesteld heeft betrekking op de ontwikkeling van een systeem voor kwaliteitszorg voor klinische co-assistentschappen in studiejaar 5 en 6. De doelstellingen zijn als volgt gespecificeerd:

- de ontwikkeling van een instrument voor een valide en betrouwbare beoordeling van co-assistentschappen;
- het creëren van voorwaarden voor de invoering van dit instrument in de onderwijspraktijk;
- het bevorderen van het gebruik van het instrument en het op grond van empirische gegevens doen van voorstellen om het onderwijs te verbeteren.

*Hoofdstuk 2* beschrijft de historische ontwikkeling van het medisch onderwijs en geeft een overzicht van vernieuwingen. Betoogd wordt dat al enige honderden jaren kritische geluiden klinken over de inrichting van het medisch onderwijs. Na een periode van betrekkelijke stabiliteit na de Tweede Wereldoorlog wordt steeds duidelijker dat studenten onvoldoende toegerust zijn voor "de gezondheidszorg van de toekomst" en alleen al daarom de opleiding aanpassing verdient. Het huidige streven is om onderwijs te realiseren dat aansluit bij de behoeften van de maatschappij en uitgaat van een studentgerichte benadering. Deze geluiden zijn niet uniek voor Nederland, maar zijn ook elders te beluisteren. Internationaal zijn er initiatieven genomen om het onderwijs te vernieuwen. De oprichting van het Network of Community-Oriented Educational Institutions for the Health Sciences (1979) is hiervan een voorbeeld (Schmidt et al., 1991). Een toonaangevend rapport waarin ideeën over de richting van mogelijke veranderingen zijn verwoord, is het Report of the Panel of the General Professional Education of the Physician and College Preparation for Medicine (Association of American Medical Colleges; GPEP-rapport, 1984). De ideeën van het Network en het GPEP-rapport zijn in vele landen actueel en vormen basis voor de herziening van veel medische curricula. Ook in Nederland heeft zich dit gemanifesteerd in aanzetten tot nieuwe ontwikkelingen binnen het onderwijs zoals studentgericht onderwijs, "community-based" onderwijs, aandacht voor psycho-sociale aspecten naast medische-technische aspecten, uitbreiding van het keuze-onderwijs en het aanbrengen van een onderwijskundig verantwoorde structuur in de opleiding.

*Hoofdstuk 3* beschrijft de opbouw van het onderwijs in de geneeskunde. Hierbij is uitgegaan van het onderwijsprogramma aan de Faculteit der Geneeskunde van de Rijksuniversiteit Limburg. Ingegaan wordt op de basisfilosofie van de Medische Faculteit Maastricht. Vervolgens wordt de opzet van de eerste vier studie jaren beschreven. Omdat deze studie betrekking heeft op de stages in studiejaar 5 en 6, is een uitvoerige beschrijving gegeven van de betreffende klinische stages.

*Hoofdstuk 4* gaat in op de voorwaarden waaraan een organisatie moet voldoen om evaluatie van stageprogramma's mogelijk te maken. De invoering van een gesystematiseerde evaluatie stelt eisen aan de context. Deze context kan men enerzijds beschrijven aan de hand van het organisatieklimaat en anderzijds aan de hand van rol-opvattingen van docenten. In de literatuur wordt een onderscheid gemaakt in rationalistische en natuurlijke organisatie modellen (Dar-

ling-Hammond & Wise en Pease, 1983). Bij een **rationalistisch** model zijn regels en bevoegdheden op centraal niveau **duidelijk** vastgelegd en is de organisatie gekenmerkt door een **planmatige** opbouw van programma's. Bij een natuurlijk model ligt de verantwoordelijkheid voor het onderwijs bij de individuele docent en is er op centraal niveau slechts een beperkte regelgeving. Bovendien maken Darling-Hammond et al. (1983) een onderscheid in vier rol-opvattingen van docenten: "teaching as labor", "teaching as craft", "teaching as art" en "teaching as profession". Het essentiële verschil tussen deze opvattingen is de mate waarin de docent autonoom kan handelen. Bij "teaching as labor" en "teaching as craft" is de autonomie van handelen vrijwel afwezig. Bij "teaching as profession" beschikt de docent over een zekere autonomie, maar binnen bepaalde grenzen. Bij "teaching as art" is de docent volledig autonoom in zijn handelen. Door de organisatie en rol-opvattingen aan elkaar te relateren kan vastgesteld worden of de context geschikt is voor een systematische evaluatiebenadering. Een **rationalistisch** organisatie-model zal samengaan met rol-opvattingen waarbij docenten minder autonoom kunnen handelen, terwijl een natuurlijk organisatie-model gekenmerkt wordt door rol-opvattingen waarbij docenten meer autonomie hebben. Een modelmatige evaluatiebenadering ligt het meest voor de hand in een **rationalistisch** model, omdat zowel de organisatie, de inhoud als de uitvoering in detail omschreven en vastgelegd zijn.

Middels een analyse van het organisatieklimaat en de docent-opvattingen die binnen de klinische stages van de Medische Faculteit Maastricht heersen werd geconstateerd dat er een duidelijke regelgeving is en taken en verantwoordelijkheden van docenten omschreven zijn. Hieruit volgt dat de situatie in Maastricht gekarakteriseerd kan worden als een **rationalistische** organisatie waarbij de taken en bevoegdheden van docenten omschreven zijn en de arts-docenten niet volledig autonoom kunnen handelen. Een modelmatige evaluatie van stages behoort dan ook tot de mogelijkheden.

Onderzoek naar de kwaliteit van onderwijs kan uitgaan van een procesgerichte of produktgerichte benadering. De operationalisering van het begrip kwaliteit verschilt al naar gelang het betrekking heeft op het onderwijsleerproces of de uitkomsten daarvan. Bij een procesgerichte benadering wordt de nadruk gelegd op de condities waaronder leren plaatsvindt. Bij een produktgerichte benadering concentreert men zich op het uiteindelijke resultaat van het leerproces. Evaluatie wordt in deze studie invulling gegeven vanuit een procesbenadering. Het belangrijkste argument hiervoor is dat vanuit een procesbenadering aanwijzingen verkregen worden over het "waarom" van de opbrengsten en daardoor specifieke informatie

wordt verkregen over de richting van gewenste veranderingen. Vanuit deze invalshoek is een instrument geconstrueerd, waarbij kwaliteit gedefinieerd is als de voorwaarden, waaraan een stage naar onderwijskundige maatstaven moet voldoen.

In *hoofdstuk 5* staat de constructie van het instrument centraal. De voorwaarden waaraan stages zouden moeten voldoen, zijn daartoe nader geconcretiseerd. Hierbij is gebruik gemaakt van studies, die zich richten op kenmerken van stages in het beroepsonderwijs en het medisch onderwijs. Een uiteindelijke beschrijving van de voorwaarden heeft plaatsgevonden door hiervan een vertaling te maken naar de huidige co-assistentschappen. Dit resulteerde in 14 inhoudelijke domeinen die kenmerkend zijn voor kwaliteit. Het betreft de volgende domeinen:

1. de voorbereiding van de studenten op de stage inzake kennis en vaardigheden;
2. de informatie die studenten ontvangen over de inhoud en opzet van de stage in de vorm van een stageboek;
3. de persoonsgebonden begeleiding die studenten ontvangen met betrekking tot hun stagewerkzaamheden;
4. de begeleiding die de student ontvangt bij het medisch-technisch handelen;
5. de aandacht die tijdens de stage gegeven wordt aan zaken die niet patiëntgebonden zijn, maar betrekking heeft op alle aspecten die te maken hebben met de patiënt en context;
6. de kwaliteit van onderwijsactiviteiten, die speciaal voor studenten georganiseerd worden en inhoudelijk expliciet zijn afgestemd op de doelgroep;
7. het aanbod van patiëntproblemen met betrekking tot het aantal en het soort patiëntproblemen waarmee de studenten in aanraking komen;
8. de kwaliteit van patiëntgebonden leersituaties die plaatsvinden in een stage, zoals verblijf op de afdeling, het meelopen met afdelingsvisites, pathologie-besprekingen en röntgenbesprekingen;
9. de kwaliteit van de polikliniek wat betreft de beschikbare ruimte om zelfstandig patiënten te kunnen onderzoeken en de leerzaamheid;
10. de kwaliteit van studiefaciliteiten die beschikbaar zijn tijdens de stage, zoals ruimtelijke voorzieningen en leermiddelen;
11. de kwaliteit van stagebeoordeling die studenten aan het einde van een stageperiode ontvangen van de opleider;

12. de gelegenheid tot zelfstudie die aan studenten gegeven wordt om naar aanleiding van praktijkervaringen zelfstudie te verrichten;
13. de tijdsbesteding tijdens de stage;
14. het algemeen oordeel van de student over de stage ten aanzien van de sfeer, de organisatie en de leerzaamheid.

Deze domeinen zijn vervolgens geoperationaliseerd in items. Het uiteindelijke meetinstrument kreeg de vorm van een vragenlijst, die studenten na afloop van elk stage dienen in te vullen. De vragenlijst is identiek voor alle stages en bestaat voor het grootste deel uit gesloten items die voorzien zijn van een vijfpunts- of Likertschaal. Voor elk item en voor elk domein wordt een gemiddelde score berekend. De domeinscores vormen de basis om beslissingen te nemen over aspecten van een stage die verandering behoeven.

*Hoofdstuk 6* gaat nader in op de wijze waarop gegevens geïnterpreteerd worden. Personen die betrokken zijn bij de inrichting, uitvoering en controle van het klinisch onderwijs ontvangen een rapport, waarin de evaluatiegegevens met de bijbehorende interpretatie zijn opgenomen. Het is belangrijk dat deze rapportage helderheid geeft over datgene dat verandering behoeft in een stage en herkenbaar is voor de gebruikers. Daarom zijn voor de rapportage twee uitgangspunten geformuleerd. Ten eerste wordt bij de rapportage uitgegaan van de onderscheiden domeinen. Dit zijn immers inhoudelijk clusters van vragen waarvan iedereen een duidelijke voorstelling heeft. Ten tweede wordt gerapporteerd op het niveau van de stage (bijvoorbeeld de stage interne geneeskunde) én het niveau van de stageplaats (bijvoorbeeld interne geneeskunde binnen het Academisch Ziekenhuis Maastricht).

Voor het geven van een betekenis aan de domeinscores wordt gebruik gemaakt van normen. Er kan een onderscheid gemaakt worden in drie typen normen: relatieve, absolute en zelfgerichte. Bij relatieve normen is de norm ontleend aan de scores van een referentiegroep. Bij absolute normen worden de scores vergeleken met een vooraf gedefinieerde standaard. Zelfgerichte normen, eveneens behorende tot de relatieve normen, worden ontleend aan voorafgaande metingen bij hetzelfde object. Elke type heeft zijn voor- en nadelen.

Bij de normen voor de stage-evaluatie is getracht om zoveel mogelijk te profiteren van de voor- en nadelen van elk type. Dit heeft geresulteerd in het gebruik van gecombineerde normen. Relatieve normen zijn het vertrekpunt voor de interpretatie. Om ervoor te zorgen dat een bepaald minimumniveau gegarandeerd blijft, geldt naast een relatieve norm een absolute norm als ondergrens. Zelfge-

richte normen, tenslotte, dienen om ontwikkelingen van een stage of stageplaats over een langere periode vast te stellen. Scores van opeenvolgende meetmomenten worden hierbij vergeleken. De relatieve beoordeling vindt plaats op drie niveaus. Ten eerste worden stages onderling vergeleken. Hierdoor krijgt men per domein een algemene kwaliteitsindicatie. Ten tweede worden verschillende stageplaatsen binnen een en dezelfde discipline vergeleken. Hierdoor krijgt men specifieke informatie over de stageplaats. Ten derde wordt een sterkte-zwakte-analyse van een stageplaats gemaakt. De gemiddelde domeinscores van een stageplaats worden hierbij onderling vergeleken waarbij de stageplaats fungeert als haar eigen referentiekader. De beschreven voorbeelden illustreren het gebruik van normen in de praktijk.

*Hoofdstuk 7* beschrijft de realisatie van de stage-evaluatie in de praktijk. Het invoeren van een systematische evaluatie van alle stages met eenzelfde instrument kan als een vernieuwing beschouwd worden. Immers het evalueren van stages gebeurde nog niet systematisch en was sterk afhankelijk van het belang dat de individuele opleider hieraan hechtte. Studies naar het verloop van innovaties maken duidelijk dat een vernieuwing meer kans van slagen heeft als aan een aantal voorwaarden is voldaan (Dalin, 1978; Fullan & Park, 1980). Binnen de instelling zijn daarom maatregelen getroffen. Er hebben gesprekken met stagecoördinatoren plaatsgevonden die er toe leidden dat iedereen in principe zijn volledige medewerking toezegde aan het doen slagen van de stage-evaluatie. Voorts zijn afspraken gemaakt over het verzamelen van gegevens, de wijze van rapportage en de verantwoordelijkheden van de betrokkenen. De instelling is erin geslaagd een klimaat te creëren waarin stage-evaluatie mogelijk is. De bruikbaarheid van evaluatiegegevens wordt geïllustreerd aan de hand van praktijkstudie 1, 2 en 3.

*Hoofdstuk 8* beschrijft twee studies naar de constructvaliditeit van het instrument. De constructvaliditeit verwijst naar de mate waarin de vragenlijst voor stage-evaluatie de bedoelde eigenschap meet. Kwaliteit is in deze studie gedefinieerd in termen van voorwaarden waaraan een stageplaats moet voldoen. Deze voorwaarden zijn omschreven in 14 benoembare domeinen en vervolgens geoperationaliseerd in items. Het begrip kwaliteit is daardoor meetbaar gemaakt. Deze domeinen zijn zowel vanuit theoretisch als pragmatisch perspectief zinvol: theoretisch refereert naar de wijze waarop het begrip kwaliteit is geoperationaliseerd op basis van literatuurstudies en de inbreng van deskundigen; pragmatisch verwijst naar de praktische bruikbaarheid in die zin dat de domeinen dusdanig zijn

gedefinieerd dat ze herkenbaar, relevant en manipuleerbaar zijn. Hieruit volgt dat er duidelijke verwachtingen geformuleerd kunnen worden over de structuur van de vragenlijst.

In studie 4 werd nagegaan in hoeverre deze op inhoudelijk en pragmatisch gespecificeerde structuur ook op empirische gronden te rechtvaardigen is. Hierbij is gebruik gemaakt van toetsende factoranalyse (LISREL), die is uitgevoerd op de data die in 1989 bij alle stages verzameld zijn ( $N=875$ ). In de analyse zijn 11 van de 14 domeinen betrokken. Drie domeinen werden niet meegenomen, daar de hiertoe behorende items open vragen waren en/of bepaalde scores ontbraken. Naast de in de praktijk gehanteerde indeling in domeinen, zijn vijf alternatieve indelingen getoetst. Voor elk model zijn de volgende indices berekend: CHI-kwadraat waarde, de grootte van de residuen (RMSR), de Goodness-of-Fit Index (GFI), de adjusted Goodness-of-Fit Index (AGFI) en de rho-waarde. Op basis hiervan is vastgesteld welke indeling op empirische gronden het beste te rechtvaardigen is. De resultaten wijzen uit dat de overeenstemming tussen het beoogde model en de data redelijk is en zeker in vergelijking met alternatieve modellen zeer acceptabel is. Besloten is om het model in voorgestelde vorm te handhaven, mede gezien de duidelijke rationale die aan het instrument ten grondslag ligt en de bruikbaarheid.

In studie 5 is onderzocht of de structuur van de vragenlijst over opeenvolgende jaren constant is. Daartoe is de invariantie tussen de correlatiematrices van de domeinen met betrekking tot 1989 ( $N=875$ ) en 1990 ( $N=851$ ) onderzocht. Geconcludeerd kan worden dat de correlaties tussen de domeinen van de vragenlijst stabiel zijn over een bepaalde periode (in dit geval twee jaar). De resultaten van beide studies vormen een duidelijke indicatie voor een redelijk goede constructvaliditeit.

*Hoofdstuk 9* doet verslag van de betrouwbaarheid van het instrument. Om hierover uitspraken te doen zijn generaliseerbaarheidsstudies uitgevoerd. Het doel hiervan is tweeledig. Ten eerste om een indicatie te krijgen van de betrouwbaarheid en daarmee indirect van de accuraatheid van de scores én van de beslissingen die genomen worden op basis van de metingen met het instrument. Ten tweede om vast te stellen hoeveel waarnemingen van studenten nodig zijn om tenminste een betrouwbaar oordeel te kunnen geven door het aantal studenten te variëren. De opzet en uitvoering van de studies is afhankelijk van datgene waarnaar gegeneraliseerd wordt, dus van het gebruiksdoel. Zoals in hoofdstuk 6 werd betoogd, kent het instrument drie gebruiksdoelen. Hieruit volgt dat drie generaliseerbaarheidsstudies zijn uitgevoerd, respectievelijk gericht op het



doen van uitspraken op niveau van de stage (studie 6), op het niveau van de stages in ziekenhuizen (studie 7) en op het niveau van de stageplaatsen, uitgaande van een sterkte-zwakte-analyse (studie 8). De beantwoording van de betrouwbaarheidsvragen heeft plaatsgevonden vanuit een relatiefgeoriënteerd en beslissingsgericht perspectief. Bij een relatief georiënteerd perspectief worden - afhankelijk van het object van meting - stages, stageplaatsen of domeinen onderling vergeleken. De betrouwbaarheid wordt dan uitgedrukt in een generaliseerbaarheidscoëfficiënt (G-coëfficiënt). Bij een beslissingsgericht perspectief wordt uitgegaan van een absoluut criterium (onder grensscore). De betrouwbaarheid wordt dan uitgedrukt in een adjusted-phi-coëfficiënt. Bovendien is de standaardmeefout berekend. De berekeningen zijn uitgevoerd met het programma GENOVA (Crick & Brennan, 1983). Belangrijke informatie ten aanzien van de onderzoeksdesigns is dat studenten en items als random variabelen opgevat worden en studenten genest zijn binnen co-assistentenschappen en gekruist met de items.

Uit de resultaten van studie 6 en 7 blijkt dat de G-coëfficiënten over het algemeen, ongeacht het aantal waarnemingen, aan de lage kant zijn. Hieruit volgt dat rangordning van stages en stageplaatsen naar één schaalwaarde verschil niet mogelijk is. De standaardmeefouten zijn daarentegen zeer acceptabel, wat een indicatie is voor de nauwkeurigheid waarmee gemeten wordt. De adjusted-phi-coëfficiënten zijn reeds bij een gering aantal waarnemingen, 25 waarnemingen per stage en 5 waarnemingen per stageplaats, zeer bevredigend. Uit de resultaten van studie 8 blijkt dat de G-coëfficiënten reeds bij 10 waarnemingen zeer acceptabel zijn, evenals de standaardmeefouten. Hieruit volgt dat bij het maken van een sterkte-zwakte-analyse de domeinen naar één schaalwaarde verschil betrouwbaar geordend kunnen worden.

Op basis van de uitkomsten van deze studies kan geconcludeerd dat het mogelijk is om met dit instrument betrouwbare uitspraken te doen op alle niveaus van interpretatie. Hieruit volgt dat de beslissingen die tot nu genomen zijn op basis van de verzamelde evaluatiegegevens verantwoord zijn. Voorts kan uit de resultaten afgeleid worden dat het niet noodzakelijk is dat alle studenten na afloop van elke stage de vragenlijst invullen. In de toekomst behoort een efficiëntere opzet van de dataverzameling tot de mogelijkheden, waarbij een aantal van 10 a-selecte studentwaarnemingen per stageplaats toereikend is voor een betrouwbare evaluatie.

In *hoofdstuk 10* volgt de discussie en een overzicht van de belangrijkste conclusies. Ingegaan wordt op de setting waarbinnen medisch onderwijs plaatsvindt en de consequenties hiervan voor

stage-evaluatie, de operationalisering van het begrip kwaliteit, het belang om alle relevante partijen te betrekken bij de stage-evaluatie, de mogelijkheden om te evalueren in een minder gestructureerde onderwijsomgeving, het achterwege blijven van directe effecten bij stage-evaluatie en het belang van een verdere professionalisering van arts-begeleiders. Op basis van de resultaten van het hier gepresenteerde onderzoek lijkt het gerechtvaardigd om te stellen dat evaluatie van klinisch onderwijs mogelijk is en bovendien dat een dergelijke benadering in de geneeskunde bruikbare informatie biedt om de kwaliteit van stages te bewaken en te bevorderen.



## SUMMARY\*

The aim of this dissertation is to describe the development of a system for quality care in clinical education, i.e. the clinical clerkships in hospitals. Chapter 1 provides a framework for this study and a (research) problem is defined. The increasing interest for quality of education in universities is highlighted. This interest is reflected and was stimulated by a number of policy measures initiated by the Dutch Ministry of Education (HOAK, 1985; HOOP, 1987) requiring universities to explicitly account for their educational quality. To provide for the necessary information evaluation systems were needed. Prior to the national interest, the University of Limburg had worked on the development of a system for quality control in education. The junior status of the university (founded in 1974) and the adoption of an innovative learning approach (problem-based learning) motivated the medical school to develop an evaluation system for quality of education. Originally the evaluation system focused on the pre-clinical years (year one through four), but after 1987 the attention also shifted to the clinical years (years five and six). In the clinical years students rotate through a number of clerkships in hospitals, family practices, and other institutions of health care. This dissertation addresses the development of such an evaluation system to assess quality of education in the clinical clerkships of year five and six. More specifically, the objectives are:

- to develop a reliable and valid instrument to assess the quality of education in clerkships;
- to establish the enabling conditions to introduce such a system;
- to promote the use of this instrument and to propose improvements in education based upon empirical data.

In a more broader perspective, *chapter 2* describes historical developments of medical education in The Netherlands. It is argued that medical education has been criticized for its structure and content for centuries. After World War II, following a relatively stable period in medical education, it becomes increasingly apparent that students are insufficiently prepared for modern health care. Education today is therefore progressively characterized by adapting to the needs of society in combination with a student centered approach. These developments are not unique for The Netherlands. The foundation of a Network of Community-Oriented Educational Institutions for the Health Sciences (Schmidt et al., 1991) is an appropriate illustration of this. Similarly, the publication of the Report of the Panel of the

\* Met dank aan Cees van der Vleuten en Jeannet Kalk.

General Professional Education of the Physician and College Preparation for Medicine (Association of American Medical Colleges; GPEP-report, 1984), declaring views on the requirements of the physician in the twenty-first century, became an authoritative document. Both the Network and the GPEP-report have stimulated innovations in medical curricula. In The Netherlands they have inspired developments such as student-centered education, community-based education, more attention for psycho-social aspects of disease, expansion of electives in education, and the establishment of an educationally founded structure in educational programs.

*Chapter 3* describes the structure of training programs in medicine in The Netherlands, using the Maastricht medical school as a frame of reference. The basic philosophy of the medical school is discussed. Subsequently, the program of the first four years are described. Given the aim of this dissertation, the program of year five and six are more elaborately discussed, including a description of the separate clinical clerkships.

*Chapter 4* addresses the organizational requirements of evaluation of clinical clerkships. The introduction of a systematic evaluation system sets specific demands to the educational context. The context can be described by the organizational climate on the one hand and by the teaching conceptions of faculty on the other hand. The literature distinguishes between rationalistic and naturalistic organizational models (Darling-Hammond, Wise & Pease, 1983). In a rationalistic model the rules and authorities are centrally defined. The organization is characterized by a coordinated structure and sequence of educational programs. In a naturalistic model individual teachers carry responsibility with limited central regulation. In relation to the teaching role Darling-Hammond et al. (1983) distinguish four conceptions, each varying in the degree of autonomy on the part of the teacher. In "teaching as labor" and "teaching as craft" hardly any autonomy exists, in "teaching as profession" the teacher has restricted autonomy within certain limits, whereas in "teaching as art" the teacher has more autonomy. Whether a systematic evaluation is achievable can be assessed by matching the organizational characteristics and the teaching conceptions. A rationalistic organization model is more likely associated with less autonomous teaching conceptions and with more central regulation. A systematic approach to evaluation is best conceivable in a rationalistic model of organization.

An analysis of the organizational climate in the clinical program and the teaching conceptions of clinical teachers involved

indicated to clear central regulations in which the tasks and authority of teachers were well defined. It was concluded that a systematic evaluation approach was potentially achievable.

The assessment of educational quality can be based on a process or outcome oriented approach. The operationalization of the concept of quality will vary depending whether the focus is on the learning process or on the outcomes of that process. In a process oriented approach the conditions of learning are emphasized. In an outcome oriented approach the final outcome of that process is essential. In this study a process oriented strategy is adopted. The main motivation for this choice is that a process oriented approach provides insight into the causes of possible outcomes and therefore allows the derivation of specific information related desirable educational changes. From this perspective, an instrument has been developed in which quality is defined as a set of educational conditions to which a clerkship has to meet.

*Chapter 5* entirely focusses on the construction of the instrument. The educational conditions to be met by the clerkships are further delineated by using studies aiming characterizing clerkships and apprenticeships in vocational training and medical education. A final definition of requisite educational conditions was established by transforming the generic literature findings to the situation of the clinical clerkships. This yielded fourteen different domains reflecting quality of education. These are:

1. preparation of students with regard to knowledge and skills;
2. received student information on content and structure of the clerkship in a formal clerkship manual;
3. personal supervision in general;
4. supervision especially of technical skill performance;
5. attention to non-patient related affairs (context of patient situation);
6. quality of educational activities directly organized for student learning;
7. patient mix in terms of problem variation and frequency;
8. quality of patient related learning situations, such as ward attendance, participation in patient rounds, pathology radiography seminars;
9. quality of the outpatients' clinic regarding available office facilities to independently examine patients;
10. quality of learning facilities available such as office facilities and learning resources;
11. quality of the clinical competence assessment given to student at the end of the clerkship;

12. opportunity for self-study as a result of patient contacts;
13. spread of activities over time;
14. general student assessment of the clerkships quality in relation to atmosphere, organization and instructiveness.

These domains are subsequently operationalized into items.

The final instrument was a uniform questionnaire which was administered at the end of all clerkships. The majority of items are in a five-point Likert type scale and a limited number required a write-in response from the students. For every item a mean score is calculated across all student responses which are subsequently aggregated to domain scores. The domain scores are used to take decisions for improving the quality of clerkships.

*Chapter 6* describes the way in which the evaluation data are interpreted. The people involved in design, realization and supervision of clinical education receive a report containing the evaluation data and their interpretation. Quite important is that the report clarifies possible suggestions for improvement amenable for change and recognizable for the target group(s). To achieve this clarity two principles are formulated. Firstly, the feedback is based upon the various dimensions, because they represent clusters of items reflecting intuitively clear conceptions of quality. Secondly, feedback is tuned to the specific clerkship (e.g. internal medicine) within location (e.g. internal medicine academic hospital Maastricht).

To interpret the domain scores standards are utilized. Three types of standards are generally employed: relative, absolute and self-directed standards. Relative standards are derived from scores of a reference group. The absolute standards are obtained by definition, usually based upon content or expertise, of required performance or scores. Self-directed standards, also a special kind of relative standards, are derived from previous measurements on the same object. Each type has its (dis)advantages, which has often resulted in combined or compromise standards. Relative standards are used in conjunction with absolute standards, the latter to guarantee a minimum requirement beyond which no compromises are made. Finally, self-directed standards are used to monitor longitudinal developments in clerkships over time.

The relative interpretation of scores here is conducted by using three different reference groups: 1) within the clerkship across domains, 2) other clerkships across all disciplines, 2) other clerkships within the same discipline. The first reference yields a profile of quality strengths and weaknesses within a clerkship, the latter two provide comparative information in relation to other clerkships. A number of practical illustrations are given in this chapter.

In *chapter 7* the practical implementation of the evaluation system is outlined. The implementation of a systematic evaluation across all clerkships can be considered as a significant innovation in the educational program, since in the existing situation evaluation occurred unsystematically and depended largely on the interest of the individual teacher. Studies of innovations indicate that success is associated with the fulfillment of a number of conditions (Dalin, 1978; Fullan & Park, 1980). A number of activities to create these conditions were therefore organized. To ensure full cooperation of the clerkship supervisors special briefing meetings were organized. Agreements were made on collection of data, ways of reporting these data and the responsibilities of those involved herein. This stimulated a climate in which clerkship evaluation became a fact. The chapter subsequently illustrates the usefulness of evaluation findings in three studies in practice (studies 1, 2, and 3).

*Chapter 8* describes two studies on the construct validation of the instrument. Construct validity here refers to the extent to which the questionnaire reflects the concept intended to be measured. By defining quality in terms of required educational conditions in fourteen different quality domains, each operationalized into sets of items reflecting that dimension, quality has become measurable. The quality domains are both theoretically relevant and practically useful. Theoretically because the domains are founded in literature and expert judgement, and pragmatically because the domains are intuitive, relevant and amenable for manipulation. Hence, clear expectations can be hypothesized on the structure of the questionnaire.

In study 4 the theoretical and pragmatic domains were tested on empirical evidence through confirmatory factor analysis (LISREL) on the 1989 data from all clerkships (N=875). Three domains were discarded from analysis because of the use of write-in questions and/or problems with missing data. In addition to the hypothesized domain division, five alternative domain models were tested. A multitude of statistical indices were used to judge the goodness-of-fit of the various models to the empirical data structure. The results indicated that the hypothesized model showed a reasonable fit and compared to the alternative models appeared best. It was concluded to maintain the suggested model.

In study 5 the reproducibility of the factor structure was investigated. The domain correlation matrices were tested on invariance across the 1989 (N=875) and 1990 (N=851) data. The results showed a stable factor structure. Both studies therefore indicated to a positive construct validity of the instrument.



*Chapter 9* investigates the reliability of the evaluation instrument by using generalizability theory. The aim was twofold. First, the aim was to obtain an estimate of the reliability, i.e. the accuracy of derived scores and of decisions taken with the instrument. Second, to assess the number of observations required to achieve sufficiently reliable scores. Various generalizability studies were undertaken following the three perspectives on decision-making that were described in chapter 6. These were studies to assess the reliability of scores for: 1) comparing clerkship quality in relation to other clerkships within the same discipline and hospital (study 6); 2) comparing clerkship quality in relation to other hospitals (study 7); 3) comparing the quality profile within a clerkship on the various domains (study 8). The interpretation of scores was both norm-referenced and mastery referenced. In a norm-referenced score interpretation scores are relatively compared across clerkships, hospitals or profiles, i.e. in relation to each other. Reliability is expressed in a generalizability coefficient (G-coefficients). In a mastery referenced perspective scores are interpreted in relation to a particular cut-off score (minimum quality) and, instead of a score reliability, a decision reliability is estimated. Reliability is expressed in an adjusted phi-coefficient. In addition, the (norm-referenced) standard error of estimate was estimated. The results of study 6 and 7 indicated fairly low G-coefficients, irrespective of the number of observations, due to relatively small clerkship variation. The standard errors of measurement, indicating the amount of noise in the measurement, were on the other hand quite acceptable, allowing significant interpretation of a single scale value. This was achieved with approximately 25 observations for clerkships and 10 observations for hospitals. The adjusted phi-coefficients were all satisfying, even with small numbers of observations (25 for clerkships and 5 for hospitals). Study 8 yielded acceptable G-coefficients and standard errors of measurement starting even at 10 observations. The general conclusion was that the instrument allows for reliable inferences at all levels of interpretation and that the assembled evaluation data in this thesis are sufficiently reproducible. It was also concluded that more efficient data-gathering is plausible by restricting the number of observations to approximately ten randomly completed student questionnaires per clerkship.

*Chapter 10* reviews the most important conclusions of this dissertation. It discusses the organizational context of medical education and its consequences for clerkship evaluation, the definition and operationalization of the quality concept, the importance of involving all parties in the evaluation, the possibilities of evaluation in less

structured situations, lack of immediate effects of evaluation, and the importance of professional development and training of clinical supervisors. The results of this thesis warrants the conclusion that clerkship evaluation is feasible and useful to improve and maintain the educational quality of clerkships.